FULLY POLYNOMIAL-TIME RANDOMIZED APPROXIMATION SCHEMES FOR GLOBAL OPTIMIZATION OF HIGH-DIMENSIONAL FOLDED CONCAVE PENALIZED GENERALIZED LINEAR MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Global solutions to high-dimensional sparse estimation problems with a folded concave penalty (FCP) have been shown to be statistically desirable but are strongly NP-hard to compute, which implies the non-existence of a pseudo-polynomial time global optimization schemes in the worst case. This paper shows that, with high probability, a global solution to the formulation for a FCP-based high-dimensional generalized linear model coincides with a stationary point characterized by the significant subspace second order necessary conditions (S^3ONC). Since the desired S^3ONC solution admits a fully polynomial-time approximation schemes (FPTAS), we thus have shown the existence of fully polynomial-time randomized approximation scheme (FPRAS) for a strongly NP-hard problem. We further demonstrate two versions of the FPRAS for generating the desired S^3ONC solutions. One follows the paradigm of an interior point trust region algorithm and the other is the well-studied local linear approximation (LLA). Our analysis thus provides new techniques for global optimization of certain NP-Hard problems and new insights on the effectiveness of LLA.

1 INTRODUCTION

This paper concerns global optimization of a folded concave penalized learning formulation for high-dimensional learning generalized linear models, which belong to statistical/machine learning problems such that the number of dimensions (or number of fitting parameters) p is (much) larger than the number of samples n. This type of problems have recently become very common in engineering and scientific applications (Fan et al., 2014a; Fan & Li, 2006). Minimal solutions to a nonconvex learning formulation have been shown effective to guarantee desirable statistical performance in order to address high dimensionality (Zhang et al., 2012). Nonetheless, generating a global solution admits no pseudo polynomial-time algorithm, unless "P = NP"; Indeed, global optimality is shown strongly NP-hard to achieve by Chen et al. (2017). In contrast to the existing pessimistic result, we derive herein a fully polynomial-time randomized approximation scheme (FPRAS) that theoretically ensure global minimality to at high probability.

More specifically, we consider a high-dimensional generalized linear model (GLM) as below. Let $X = (x_1, ..., x_n)^\intercal$ be the $n \times p$ design matrix with $x_i = (x_{i1}, ..., x_{ip})^\intercal,\ i = 1, ..., n$, and $Y = (y_1, ..., y_n)^\intercal$ be the n-dimensional response vector. We will assume the design matrix X is fixed, while the mean of the response is given by $E[y_i] = \psi'(x_i^\top \beta^{true})$ for some known link function $\psi:\ \Theta \to \Re$, where $\Theta \subseteq \Re$ and $\beta^{true} = (\beta_1^{true}, ..., \beta_p^{true})$ is the unknown vector of true parameters of the model. GLMs are an extension of linear regression models, allowing for a flexible approach to model estimation. The high-dimensional regression problem is to estimate β^{true} given knowledge of X, Y, and ψ in the undesirable scenario where $p \gg n > 0$. To that end, traditional statistical learning schemes often resort to the following formulation:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \ell(y_i, x_i, \beta) = \frac{1}{n}\sum_{i=1}^{n}[\psi(x_i^\intercal \beta) - y_i x_i^\intercal \beta]. \tag{1}$$

which, according to traditional statistical theories, would result in overfitting in general under the high-dimensional setting.

To resolve overfitting, modern statistical theories favor a modified formulation as below:

$$\min_{\beta} \left[\mathcal{Q}(\beta) := \mathcal{L}(\beta) + \sum_{j=1}^{p} P_\lambda(|\beta_j|) \right], \tag{2}$$

where $P_\lambda(|\cdot|)$ is sparsity-inducing regularization term that penalizes any nonzero dimensions in the minimizer, and $\lambda > 0$ is a tuning parameter. Under the assumption that the true regression parameter β^{true} is sparse, a global optimizer to equation 2 has been shown effective to address overfitting for many choices of specific regularization functions P_λ. Indeed, one of the most successful choice of P_λ is the much studied Lasso-based regularized (Tibshirani, 1996), aka, the ℓ_1(-norm) penalty, which has been demonstrated to entail desirable statistical properties (Bickel et al., 2009; Negahban et al., 2012). Another admirable property of the Lasso is that, especially when applied to least squares linear regression, it yields an extremely tractable problem via a variety of algorithms (Friedman et al., 2008; 2010). However, per Zhao & Yu (2006); Fan & Li (2001), Lasso is not selection consistent without a strong irrepresentable condition and may sometimes introduce non-trivial estimation bias.

As a popular alternative to Lasso, the folded concave penalty (FCP) is first introduced by Fan & Li (2001). There are mainstream examples of FCP functions, including the SCAD by Fan & Li (2001) and MCP by Zhang et al. (2010). This paper will focus on MCP, defined as $P_\lambda(|t|) = \int_0^{|t|} \frac{(a\lambda - s)_+}{a} ds$ for some fixed parameter $a > 0$. In contrast to the Lasso, the FCP achieves variable selection consistency non-contingent on an irrepresentable condition and is demonstrated to be unbiased (Fan & Li, 2001). Furthermore, Zhang et al. (2012) demonstrated that the global solution the FCP-regularized formulation leads to desirable recovery of the oracle solution.

Nonetheless, FCP problems are significantly harder to solve than Lasso, the new penalty term moves the problem outside the realm of convex optimization, Chen et al. (2017) even showed that any estimation problem with convex loss and folded concave regularization to be strongly NP-hard, ruling out the possibility of a pseudo-polynomial-time global optimization algorithm. Liu et al. (2016) were seemingly the first to propose a global approach to the problem called MIPGO which reformulates the problem into a mixed integer program. Yet, the worst-case complexity of MIPGO is in exponential time.

Alternatively, recent literature focuses on local algorithms for the FCP-regularized learning problems. The local quadratic approximation algorithm by Fan & Li (2001) is an example of a majorization minimization algorithm, an approach which is also related to the local linear approximation (LLA) algorithm proposed by Zou & Li (2008). LLA was further explored by Fan et al. (2014b) showing the oracle property can be obtained with high probability despite the local approach. Mazumder et al. (2011); Fan & Lv (2011) demonstrate coordinate optimization approaches for FCP while Wang et al. (2014) used an approximate regularization path-following algorithm to obtain the optimal convergence rate to statistically desirable local solution. Wang et al. (2013) analyzed the CCCP algorithm and showed under certain conditions that it asymptotically finds the oracle estimator. Liu et al. (2017) took an algorithm agnostic approach by analyzing local solutions satisfying second order KKT conditions and showed desirable statistical properties like recovering the oracle solution and sparisty. The above approaches have mainly focused on linear regression, a special case of GLM where ψ is specifically the identity function. For analyses which encompass GLM's with FCP regularizers, Fan & Lv (2011) showed that GLM's, even in ultra high dimensional variable selection problems, have oracle properties when using FCP regularization and demonstrated a coordinate optimization algorithm for finding local solutions. In the area of M-estimators, which is a further generalization of our estimation method beyond even GLMs, Loh & Wainwright (2013); Loh et al. (2017a) showed that under certain conditions all local solutions must be within statistical precision of the true parameter and its support while Loh et al. (2017b) demonstrate a two-step algorithm involving composite gradient descent to find a local solution.

From the numerous results pertaining local solution schemes above, our research question is why local solutions are repetitively successful. In other words, are there certain geometric properties of the learning formulation equation 2 with FCP that allow all local schemes to perform well independent of the specific designs of the algorithmic procedures? Our answer to this question is affirmative;

we show herein that all local solutions within an efficiently achievable sub-level set are actually globally optimal. Those local solutions are characterized by the significant subspace second-order necessary conditions (S^{3}ONC) and are provably computable within pseudo-polynomial time. The S^{3}ONC are weaker conditions than the standard second-order KKT conditions. As per this result, all S^{3}ONC-guaranteeing algorithms (which include a large spectrum of local algorithms) belong to the class of FPRAS's for global optimization of the strongly NP-hard FCP-based learning problem. We subsequently develop theories for two specific algorithms of this type: one gradient-based method and the other is the same as the LLA.

It is worth noting that Zhang et al. (2010) provides conditions to establish the uniqueness of local solutions to FCP-based learning. When local solutions are unique, then any local optimization algorithms would ensure global optimality. However, a few critical assumptions are necessary to achieve the uniqueness result and, furthermore, many report numerical experiments, e.g., those in Fan et al. (2014b); Liu et al. (2017; 2016); Fan & Li (2001) indicate the non-uniqueness of local solutions, instead. In contrast, our results in this paper imposes only standard assumptions commonly shared by a flexible set of high-dimensional GLMs and are applicable even if the local solutions are non-unique. To our knowledge, this is the first geometric proof that global solutions coincides with pseudo-polynomial-time computable local solutions in an FCP-based regression formulation with high probability, despite that local solution are not necessarily unique. The resulting algorithms are the first few FPRAS's to this problem.

The rest of this paper is organized as follows. Section 2 goes through specific problem assumptions and explains the S^{3}ONC. Section 3 contains our main result for global optimality and uses it to make additional claims for LLA. Section 4 contains numerical results to verify our theoretical results. Section 5 provides concluding remarks.

In this paper we will use $\|\cdot\|_0$ to denote the number of nonzero entries, $|\cdot|$ to denote the ℓ_1-norm if the argument is a vector or cardinality if the argument is a set, $\|\cdot\|$ to denote the ℓ_2-norm, $\|\cdot\|_{max}$ to denote the maximum norm and $\|\cdot\|_{min}$ to denote the absolute value of the entry with the smallest magnitude. $(\cdot)_+$ is used equivalently to $\max(0,\cdot)$. For any vector v, $v_{\mathcal{Q}}$ is intended as $(v_j : j \in \mathcal{Q})$. For any set $\mathcal{Q}$, we denote the complement as $\mathcal{Q}^c$. In particular, let S be the true support set, that is, $\mathcal{S} := \{j : \beta_j^{true} \neq 0\}$ and its complement is $\mathcal{S}^c$. We occasionally use the term the "oracle solution" to refer to the solution β^{oracle} defined as $\beta^{oracle} \in \underset{\beta:\, \beta_j=0,\, \forall j \notin \mathcal{S}}{\arg\min} \mathcal{L}(\beta)$. The oracle solution is a hypothetical assumes the prior knowledge on the true support set S and thus can be considered a theoretical benchmark.

2 Setups, preliminaries, and assumptions

2.1 Setups and assumptions

Our analysis will focus on GLMs that have a fixed design matrix and satisfy the following assumptions:

(A1) Assume that

  (i) $b_u \geq \psi''(x_i^\intercal \beta) \geq b_l > 0$ for all $x_i^\intercal \beta \in \Theta$;

  (ii) There exists $K > 0$ such that the design matrix satisfies $\frac{1}{n}\|X_j\|^2 < K$ for all $j \in [p]$. Let the tuning parameter a in P_λ satisfy $K < (b_u a)^{-1}$.

(A2) The vector of residuals $W \in \Re^n$ is subgaussian(σ) which means it satisfies that $P[|\langle W, v\rangle| \geq t] \leq 2\exp(-t^2/2\sigma^2)$, for all $v : \|v\| = 1$ and any $t > 0$;

(A3) There exists a sequence $\{r_d \geq 0 : d = 1, 2, ..., p\}$ such that the following are satisfied:

  (i) For any $d_1, d_2 : 1 \leq d_1 \leq d_2 \leq p$, we have $r_{d_1} \geq r_{d_2}$;

  (ii) There exists some $\tilde{p}^* : 2|\mathcal{S}| \leq \tilde{p}^* \leq p$ such that $r_{\tilde{p}^*} > 0$;

  (iii) For all $d : 1 \leq d \leq p$, it holds that $n^{-1}\|X\beta\|^2 \geq r_d\|\beta\|^2$ for any $\beta \in \Re^p : \|\beta\|_0 \leq d$.

Remark 1. *Part (i) of (A1) states that our link function is both strongly convex and continuously differentiable; that is, the gradient gradient being Lipschitz continuous. Several distributions found in traditional GLM problems satisfy this constraint including normal, gamma, Poisson, categorical* *and multinomial distributions, although in some cases the mild assumption on the boundedness of Θ has to be made additionally. Even though the original domain of the link function can be unbounded, one may still consider its bounded subset given that it contains the vector of true parameters. Part (ii) of (A1) can be assumed without loss of generality by normalizing the design matrix columns.*

Remark 2. *(A2) is a common assumption Negahban et al. (2012) Wang et al. (2013) which applies to a variety of traditional GLM problems including normal, categorical and multinomial distributions.*

Remark 3. *Assumption (A3) can be understood to be a lower bound on the eigenvalues for principal sub-matrices of $X^\intercal X$ of dimension $d \times d$ for all $d \in [p]$. For every $d: \ d \leq \tilde{p}^*$, the lower bounds are positive, meaning that the smallest eigenvalues of the $d \times d$ principal sub-matrices are assumed positive.*

According to Liu et al. (2017), Assumption (A3), for certain parameters, is provably a weaker condition than the restricted eigenvalue (RE) condition, as defined in Definition 1 below and first introduced by Bickel et al. (2009) as a plausible assumption to allow for the desired recovery quality of Lasso. The RE is a common assumption in the high-dimensional learning literature, such as Zhang et al. (2014) and Fan et al. (2014b).

Definition 1. *(RE condition Zhang et al. (2010)) The matrix $X \in \Re^{n\times p}$ is said to satisfy the RE condition if, for some $r_e > 0$, it holds that $\frac{1}{n}\|X\delta\|^2 \geq r_e \|\delta\|^2$ for all $\beta \in \bigcup_{|\hat{S}|=s} \mathbb{C}(\hat{S})$ where $\mathbb{C}(\hat{S}) := \{\delta := (\delta_i) \in \Re^p : |\delta_{\hat{S}^c}| \leq 3|\delta_{\hat{S}}|\}, \delta_{\hat{S}^c} := (\delta_j : j \in \hat{S}^c)$, and $\delta_{\hat{S}} := (\delta_j : j \in \hat{S})$. Furthermore, the largest possible r_e is said to be the restricted eigenvalue constant of X.*

Random design matrices generated following subgaussian distributions under some independence assumptions have been shown to satisfy the RE condition with high probability by Zhou (2009). Thus (A3) is also satisfied with high probability under the same setting.

2.2 Preliminaries on S^{3}ONC

Our results focus on the S^{3}ONC solutions, which has been formerly introduced by Liu et al. (2017) in the special case of high-dimensional linear regression as a relaxation of the standard second-order KKT conditions. The definition of S^{3}ONC depends on the notion of first order necessary conditions (FONC) as below.

Definition 2 (FONC)**.** *A solution β^* satisfies the first order necessary conditions (FONC) if*

$$\exists \mathcal{D}(\beta^*) \in 1/n \sum_{i=1}^{n} [\psi'(x_i^\intercal \beta^*) - y_i]\, x_i + \left(P'_\lambda(|\beta_j^*|)\partial(|\beta_j^*|), 1 \leq j \leq p\right) \quad s.t.\ \mathcal{D}(\beta^*) = 0 \tag{3}$$

where $\partial(|\cdot|)$ denotes the subdifferential of $|\cdot|$.

Definition 3 (S^{3}ONC)**.** *A solution β^* satisfies the significant subspace second-order necessary condition (S^{3}ONC) if it satisfies FONC and for all $j \in \{j : \beta_j^* \neq 0\}$,*

$$\left.\frac{\partial^2 \mathcal{Q}(\beta)}{(\partial \beta_j)^2}\right|_{\beta=\beta^*} \geq 0 \tag{4}$$

if the second derivative exists.

Remark 4. *The S^{3}ONC can be intuited as the second order necessary condition applied only to the dimensions where $\beta_j \neq 0$, i.e., the significant dimensions. Since the S^{3}ONC is weaker than the standard second-order KKT conditions, any algorithm that guarantees the second-order KKT conditions can be used to obtain an S^{3}ONC solution, by requiring a more stringent optimality condition, may be slower than necessary. One specifically S^{3}ONC guaranteeing approach, presented in Liu & Ye (2019), utilizes an interior point trust region algorithm in order to guarantee an $S3ONC$ solution in polynomial time. This is the scheme which will be used later in Section 4.*

3 MAIN RESULTS

We now present our theoretical results for global optimization of FCP penalized GLMs. All proofs can be found in the appendix. We will make use of a short-hand notation:

$$\beta^{Lasso} \in \arg\min \mathcal{L}(\beta) + \lambda|\beta|. \tag{5}$$

Theorem 1. *Suppose assumptions (A1), (A2), and (A3) with any $\tilde{p}^* : \tilde{p}^* \geq 2|\mathcal{S}|$. Let β^* be an arbitrary S^3ONC solution to equation 2 with P_λ specified as the MCP. Assume that $\mathcal{Q}(\beta^*) \leq \mathcal{Q}(\beta^{true}) + \Gamma$ for an arbitrary $\Gamma \geq 0$. (i) Let the sub-optimality gap satisfy $\Gamma < P_\lambda(a\lambda) - \frac{\sigma^2}{b_l n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right)$; (ii) choose $P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1 + 2\sqrt{t'} + 2t') + \frac{\frac{\sigma^2}{n}|\mathcal{S}|(1+2\sqrt{t'}+2t')+\Gamma b_l}{b_l(\tilde{p}^*-2|\mathcal{S}|+1)}$ and (iii) assume that the minimal signal strength satisfy*

$$\left\|\beta_{\mathcal{S}}^{true}\right\|_{\min} > \sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) + \frac{8}{r_{\tilde{p}} b_l} \min\left\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathcal{S}|,\ P_\lambda(a\lambda)|\mathcal{S}| + \Gamma\right\}}$$

then the following two statements hold

(a) β^* is an oracle solution with probability at least $1 - \exp(-t + \tilde{p}^* \ln(\frac{pe}{\tilde{p}^*})) - \exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}$.

(b) β^* is both an oracle solution and an globally optimal solution to equation 2 with probability $1 - 2\exp(-t + \tilde{p}^* \ln(\frac{pe}{\tilde{p}^*})) - 2\exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}$.

Remark 5. *Theorem 1 (especially in the second statement) is perhaps the first result that establishes a set of conditions for any S^3ONC solution to be globally optimal with high probability. Further, this result is algorithm independent which allows for greater flexibility compared to most existing results as in Loh et al. (2017b) and Fan & Lv (2011) which rely on a specific algorithm choice.*

Remark 6. *The second part follows quite easily from the first due to the uniqueness of β^{oracle} as well as the fact that β^{opt} must also be an S^3ONC solution. Thus by applying the first part of the Theorem to β^{opt} we are able to show that both our arbitrary β^* and β^{opt} coincide with the unique β^{oracle}.*

Remark 7. *The above constraints on Γ, $P_\lambda(a\lambda)$ and $\|\beta_{\mathcal{S}}^{true}\|_{min}$ may initially seem disparate but can all be converted to constraints on the sample size n as is shown in Corollary 1 below.*

Corollary 1. *Assume $\ln p \geq 1$, $b_l \leq 1$, and $s \geq 1$. Let β^* be an S^3ONC solution to equation 2. Let assumptions (A1), (A2), and the RE condition as defined in Definition 1 hold. Assume that $\mathcal{Q}(\beta^*) \leq \mathcal{Q}(\beta^{Lasso})$ almost surely, where β^{Lasso} is the optimal solution to the Lasso problem with penalty coefficient $\lambda^{Lasso} = \sigma\sqrt{\frac{lnp}{n^{1-\gamma/2}}}$ where $\gamma \in [0,1]$ is an arbitrary scalar. Let $\lambda = \frac{\sigma}{r_e}\sqrt{\frac{\ln p}{n^{\gamma/2}}}$ and $a \in [.8, 1)$. There exist problem independent constants $C_1 > 0$,$C_2 > 0$ and $C_2 > 0$ such that if*

$$n > \max\left\{\frac{C_1}{b_l}, \left[C_2\frac{s}{b_l}\right]^{\frac{2}{1-\gamma}}, \left[C_3\frac{s\sigma^2 \ln p}{\|\beta_{\mathcal{S}}^{true}\|_{min}^2 b_l^2 r_e^4}\right]^{2/\gamma}\right\} \tag{6}$$

Then β^* is the global solution to 2 with probability at least $1 - C_4\exp(-C_5 s n^{\gamma/2}\ln p) - C_6\exp(-C_7 b_u n^{\gamma/2}\ln(p))$ for problem independent constants C_4, C_5, C_6 and C_7

Remark 8. *Corollary 1 indicates that for $\gamma > 0$, the global optimal solution coincides with computable S^3ONC solution with overwhelming probability given that the sample size meets certain requirements. It should specifically be noted that the relationship between n and p require only $\frac{\ln p}{n^{\gamma/2}} = O(1)$, which ensures the applicability to the high-dimensional setting even if $n \ll p$.*

Remark 9. *Liu & Ye (2019) has derived a gradient-based algorithm that provably ensures an S^3ONC solution at pseudo-polynomial-time complexity. When n is properly large, this pseudo-polynomial-time algorithm enables a straightforward design of an FPRAS for generating the global optimal solution as follows.*

FPRAS: A pseudo-polynomial-time algorithm that generates global optimal at high probability

Step 1. Initialize the algorithm with β^{Lasso} by solving equation 5

Step 2. Invoke the gradient-based algorithm (Algorithm 1 in Liu & Ye (2019)) with initializer β^{Lasso}.

Remark 10. *The stipulation that $\mathcal{Q}(\beta^*) \leq \mathcal{Q}(\beta^{Lasso})$ can generally be obtained by initializing any S^3ONC guaranteeing algorithm with β^{Lasso} in a similar fashion to Fan et al. (2014b) for LLA. The FPRAS above follows the same initialization scheme.*

Remark 11. *The above specification of values for a, λ and λ^{Lasso} can be thought of as examples rather than strict requirements. A closer examination of the proof for Corollary 1 will reveal that the values for λ and λ^{Lasso} can be chosen in a much more flexible fashion, though the corresponding values of C_1 through C_7 may be different for different combinations of λ and λ^{Lasso}.*

The techniques used in the proof of Theorem 1 can be used to provide insights into other optimization schemes. As an example, we can apply the same analysis to the state-of-the-art FCP-based algorithm, LLA, using the framework in Fan et al. (2014b) as a starting point.

LLA: local linear approximation.

Step 1. Set $k = 0$. Initialize the algorithm with $\beta^0 = \beta^{Lasso}$, where β^{Lasso} is generated by solving equation 5. Let N be the maximal iteration number.

Step 2. For all $k = 1, ..., N$, solve the following convex program to generate β^{k+1}:

$$\beta^{k+1} \in \arg\min \mathcal{L}(\beta) + \sum_{j \in [p]} P'_\lambda(|\beta_j^k|) \cdot |\beta_j|,$$

where P'_λ is the first derivative of P_λ. Let $k := k + 1$.

We can show that in fact the LLA is another FPRAS that achieves the global optimal solution. The proof of this can be found in the appendix.

Corollary 2. *For problem equation 2. If $\|\beta_{\mathcal{S}}^{true}\|_{min} > (a + 1)\lambda$, $\lambda > \max\{\frac{3\lambda^{Lasso}s^{1/2}}{b_l r_e}, \frac{4\sigma\sqrt{s+2\sqrt{st_1}+2t_1}}{b_l(an/b_u)^{1/2}}, \frac{2\sigma\sqrt{s+2\sqrt{st_2}+2t_2}}{b_l\sqrt{nr_e}}\}$ and the RE condition in Definition 1 holds, the following holds.*

(a) The LLA algorithm initialized with β^{Lasso} converges to the oracle solution in two iterations with probability $1 - \phi_0 - \phi_1 - \phi_2$, where

$$\phi_0 := P(\left\|\beta^{Lasso} - \beta^{true}\right\|_{max} > \lambda) \leq 2p\exp(\tfrac{-(\lambda^{Lasso})^2 nb_u a}{8\sigma^2}),$$

$$\phi_1 := P(\left\|\nabla_{S^c_{\tilde{p}}}\ell_n(\beta^{oracle})\right\|_{max} \geq \lambda) \leq (\tfrac{pe}{s})^s\exp(-t_1) + 2\exp(\tfrac{-\lambda^2 ab_u n}{8\sigma^2}),$$

$$\phi_2 := P(\left\|\beta_{\mathcal{S}}^{oracle}\right\|_{\min} \leq a\lambda) \leq (\tfrac{pe}{s})^s\exp(-t_2),$$

(b) If in addition (A1) and (A2) holds, while the parameters of (a, λ) satisfy that $P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1 + 2\sqrt{t_4} + 2t_4) + \frac{\frac{\sigma^2}{n}|\mathcal{S}|(1+2\sqrt{t_4}+2t_4)b_l}{b_l(\tilde{p}^*-2|\mathcal{S}|+1)}$ and the and $P_\lambda(a\lambda) > \frac{\sigma^2}{b_l n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t_3} + 2t_3\right)$ and let the minimal signal strength satisfy $\|\beta_{\mathcal{S}}^{true}\|_{\min} > \sqrt{\frac{8\sigma^2}{r_{\tilde{p}}b_l^2 n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t_3} + 2t_3\right) + \frac{8}{r_{\tilde{p}}b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathcal{S}|, P_\lambda(a\lambda)|\mathcal{S}|\}}$ then the LLA algorithm initialized by β^{Lasso} converges to the global solution in two iterations with probability at least $1 - \phi_0 - \phi_1 - \phi_2 - \phi_3$

where

$$\begin{aligned}\phi_3 := P(B^{oracle} \neq B^{opt}) \leq \exp(-t_3 + \tilde{p}^*\ln(\frac{pe}{\tilde{p}^*})) \\ + \exp(-(\tilde{p}^* + 1)(t_4 - \ln p)) \cdot \frac{1 - \exp(-(p - \tilde{p}^*)(t_4 - \ln p))}{1 - \exp(-t_4 + \ln p)},\end{aligned} \quad (7)$$

and a_0, a_1 are defined as in Fan et al. (2014b) and $t_1, t_2, t_3, t_4 > 0$ are arbitrary constants.

Remark 12. *Since each iteration of the LLA solves a convex program, which can be done within polynomial-time. When n is properly large, the above theorem then indicates that the LLA is another FPRAS in globally optimizing the FCP-based nonconvex formulation.*

4 Numerical experiments

4.1 Experimental setup

We focus our tests on sparse logistic regression. Our problem and data are implemented in a similar way as Fan et al. (2014b). We construct β^{true} as below: Firstly, β_S^{true} is constructed randomly by choosing 10 elements of β and choosing the magnitude of each to be a uniform value within $[1, 2]$. Each value is chosen to be negative with probability .5. Then, the remaining entries $\beta_{S^c}^{true}$ are set to be 0. The design matrix $X \in \Re^{n\times p}$ is constructed by generating n iterations of $x_i \sim N_p(0, \Sigma)$ where $\Sigma = (.5^{|j-j'|})_{p\times p}$. We then generate Y using a Bernoulli distribution where $P(y_i = 1) = (1 + e^{-x_i^\intercal \beta^{true}})^{-1}$. Utilizing this process, we generate two sets of data, both with 100 samples. One set is for training the model, and the other is the test set for out-of-sample tests. We repeat the above process for 100 times to generate 100 training-and-test instances, each with 100 samples.

We train a logistic regression model by invoking Algorithm 1 in solving equation 2 with FCP for S^{3}ONC solutions initialized with Lasso. For comparison, we also involve Lasso solutions generated by the global minimizer to equation 5 and an estimator generated by solving equation 2 when P_λ is substantiated by an ℓ_2 penalty. The tuning parameters λ and a (if applicable) of the penalties for the estimators are obtained by cross validation following Fan et al. (2014b). The MCP classifier is solved using the FPRAS from Liu & Ye (2019) implemented in Python 3. The Lasso and ℓ_2 classifiers are solved using the *scikit learn* python library.

We compare the above estimators on the statistical performance. We use both ℓ_1 loss: $|\beta^* - \beta^{true}|$ and ℓ_2 loss: $\|\beta^* - \beta^{true}\|$.

Finally, we try to ascertain whether our FCP classifier, obtained using S^3ONC methods is actually the global optimal solution. We do this by taking each element of the FCP classifier and perturbing each element. Each element's perturbation is independent and generate by a $N(0, 1/p^{1/2})$)-random variable. We then check if this perturbed classifier has better FCP penalized performance on the training data than the FCP classifier. If not, we repeat until either a better solution is found, or until 2000 perturbations have been tried.

4.2 Numerical results

Table 1: Statistical performance of the four classifiers.

		$n = 100,\ p = 1000$		$n = 100,\ p = 1500$		$n = 100,\ p = 2000$	
Classifier	**Measure**	**Mean**	**Std. dev**	**Mean**	**Std. dev**	**Mean**	**Std. dev**
FCP	ℓ_1 loss	13.909907	1.471911	14.818059	1.698191	14.506226	1.480686
	ℓ_2 loss	4.108019	0.320061	4.304993	0.374453	4.489184	0.399441
Lasso	ℓ_1 loss	15.015975	1.039529	15.882654	1.29422	17.079414	1.545309
	ℓ_2 loss	4.3255	0.25996	4.397969	0.326336	4.433467	0.362707
ℓ_2 penalty	ℓ_1 loss	22.211963	0.791955	26.026067	0.966091	28.485075	0.993699
	ℓ_2 loss	4.734209	0.241683	4.738025	0.296726	4.755959	0.296746

Table 2: Percent of time FCP beat all perturbations

	$n = 100$ $p = 500$	$n = 100$ $n = 1000$	$n = 100$ $n = 1500$	$n = 100$ $n = 2000$
% Best FCP	100%	100%	100%	100%

Table 1 shows the numerical results for the statistical performance measurements. We show the two performance measures for each of the three classifiers for tphree different problem types.

As expected, the FCP classifier generally outperformed the lasso and ℓ_2 classifiers. The margins are fairly thin between FCP and lasso, especially compared to the standard deviation. Other values of n and p were tried but the results generally followed the same pattern.

Table 2 contains the numbers from optimality analysis. This technique did not yield a single perturbed solution that could beat the FCP classifier obtained from the FPTRAS in any of our thousands of iterations.

As a result we tentatively conclude that our numerical results align with our theoretical results though further testing of the global optimality probability would be valuable.

5 CONCLUSIONS

This paper investigates both the theoretical and empirical performance of pseudo-polynomial time algorithms on FCP regularized GLMs. Despite such a problem being strongly NP-Hard, we have shown two FPTRAS that achieve global optimality. To our knowledge this is the first probability bound for pseudo-polynomial time global optimization of FCP regularized GLMs. Further, the same technique can be used to extend other results in order to obtain global optimization bounds for a wide variety of problems.

Though this paper focuses on GLMs, further exploration will focus on the question whether similar results can be found for more general problem classes under weaker assumptions. High-dimensional M-estimation problems could potentially be a future avenue of investigation.

REFERENCES

Peter J Bickel, Yaacov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Yichen Chen, Dongdong Ge, Mengdi Wang, Zizhuo Wang, Yinyu Ye, and Hao Yin. Strong np-hardness for sparse optimization with concave penalty functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 740–747. JMLR. org, 2017.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Runze Li. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*, 2006.

Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1 (2):293–314, 2014a.

Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42(3):819, 2014b.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Hongcheng Liu and Yinyu Ye. High-dimensional learning under approximate sparsity: A unifying framework for nonsmooth learning and regularized neural networks. *arXiv preprint arXiv:1903.00616*, 2019.

Hongcheng Liu, Tao Yao, and Runze Li. Global solutions to folded concave penalized nonconvex learning. *Annals of statistics*, 44(2):629, 2016.

Hongcheng Liu, Tao Yao, Runze Li, and Yinyu Ye. Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory for local solutions. *Mathematical programming*, 166(1-2):207–240, 2017.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pp. 476–484, 2013.

Po-Ling Loh, Martin J Wainwright, et al. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017a.

Po-Ling Loh et al. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 2017b.

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Lan Wang, Yongdai Kim, and Runze Li. Calibrating non-convex penalized regression in ultra-high dimension. *Annals of statistics*, 41(5):2505, 2013.

Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.

Cun-Hui Zhang, Tong Zhang, et al. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pp. 921–948, 2014.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.

A Appendix

The Appendix is organized as below: Section A.1 presents the proofs for the main results, Sections A.2 and A.3 present central lemmata to be useful in Section A.1.

A.1 PROOF OF MAIN RESULTS

A useful relationship in our proofs is that, for an S^{3}ONC solution β^* within $\{\beta^* : \mathcal{Q}(\beta^*) \le \mathcal{Q}(\beta^{true}) + \Gamma\}$ for any $\Gamma \ge 0$, we have the following useful inequality under Assumption (A1):

$$\frac{b_l}{2n} \|X\delta^*\|^2 - \frac{1}{n} W^\intercal X \delta^* + \sum_{j\in\mathcal{S}} P_\lambda(|\beta_j^*|) \le \sum_{j\in\mathcal{S}} P_\lambda(|\beta_j^{true}|) + \Gamma, \tag{8}$$

where $\delta^* = \beta^* - \beta^{true}$. This is obtained by invoking the strong convexity of ψ, which leads to $\psi(x_i^\intercal \beta^*) \ge \psi(x_i^\intercal \beta^{true}) + \psi'(x_i^\intercal \beta^{true})(x_i^\intercal \beta^* - x_i^\intercal \beta^{true}) + 0.5 \cdot b_l (x_i^\intercal \beta^* - x_i^\intercal \beta^{true})^2$.

Proof of Theorem 1. First, given our assumption that (A1) holds, that (i) $\tilde{p}^* \ge 2|\mathcal{S}|$, (ii) β^* is S^3ONC satisfying $\mathcal{Q}(\beta^*) \le \mathcal{Q}(\beta^{true}) + \Gamma$ for some $\Gamma \ge 0$, and (iii) $P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1 + 2\sqrt{t'} + 2t') + \frac{\frac{\sigma^2}{n}|\mathcal{S}|(1+2\sqrt{t'}+2t')+\Gamma b_l}{b_l(\tilde{p}^*+1-2|\mathcal{S}|)}$, we can apply Lemma 5 with $\tilde{p} = \tilde{p}^*$. This means that $\|\beta^* - \beta^{true}\| \le \tilde{p}^*$ with probability at least $1 - \exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}$. From this, given the additional assumption that (A3) holds, we can apply the second part of Lemma 4 with $\tilde{p} = \tilde{p}^*$ to get that for any $t > 0$, $\frac{1}{n}\|X(\beta^* - \beta^{true})\|^2 \le \frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) + \frac{8}{b_l}\min\{\lambda^2(|\mathcal{S}| - \|\beta^*\|_0) r_{\tilde{p}^*}^{-1}, P_\lambda(a\lambda)\dot{(}|\mathcal{S}| - \|\beta^*\|_0) + \Gamma\}$ holds with probability at least $1 - \exp(-t + \tilde{p}^* \ln(\frac{pe}{\tilde{p}^*}))$. Given that for 2 arbitrary sets A and B

$$\begin{aligned} P(A \cap B) = P(B)P(A|B) &= (1 - P(B^c))(1 - P(A^c|B)) \\ &= 1 - P(A^c|B) - P(B^c) + P(B^c)P(A^c|B) \\ &= 1 - P(A^c|B) - P(B^c)(1 - P(A^c|B)) \ge 1 - P(A^c|B) - P(B^c) \end{aligned} \tag{9}$$

Therefor they hold simultaneous with probability at least $1 - \exp(-t + \tilde{p}\ln(\frac{pe}{\tilde{p}^*})) - \exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}$.

The same sequence of arguments can be used to show that β^{opt} also satisfies $\|\beta^{opt} - \beta^{true}\| \le \tilde{p}^*$ and $\frac{1}{n}\|X(\beta^{opt} - \beta^{true})\|^2 \le \frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) + \frac{8}{b_l}\min\{\lambda^2(|\mathcal{S}| - \|\beta^*\|_0) r_{\tilde{p}^*}^{-1}, P_\lambda(a\lambda)\dot{(}|\mathcal{S}| - \|\beta^*\|_0) + \Gamma\}$ with the same probability. Using again the union bound and DeMorgan's law, we say β^* and β^{opt} satisfy the above conditions simultaneously with probability $1 - 2\exp(-t + \tilde{p}^* \ln(\frac{pe}{\tilde{p}^*})) - 2\exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}$. With this, our Γ assumption and our minimal signal strength assumption, we can apply Lemma 6 to show that $\beta^* = \beta^{opt}$ with probability at least $1 - 2\exp(-t + \tilde{p}^* \ln(\frac{pe}{\tilde{p}^*})) - 2\exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t'-\ln p))}{1-\exp(-t'+\ln p)}$. □

Proof of Corollary 1. First we need to bound Γ. In order to do this we use the lasso problem $\mathcal{Q}^{lasso}(\beta) = \sum_{i\in\mathcal{N}} \ell(\beta, x_i, y_i) + \sum_{j\in\mathcal{P}} \lambda^{lasso}|\beta_j|$ as well as the concavity of MCP over positive values to get the following 2 inequalities

$$\begin{aligned} &\mathcal{Q}^{lasso}(\beta^{lasso}) \le \mathcal{Q}^{lasso}(\beta^{true}) \\ &\sum_{i\in\mathcal{N}} \ell(\beta_j^{lasso}, x_i, y_i) - \ell(\beta_j^{true}, x_i, y_i) \\ &\le \sum_{j\in\mathcal{P}} \lambda^{lasso}(|\beta_j^{true}| - |\beta_j^{lasso}|) \le \sum_{j\in\mathcal{P}} \lambda^{lasso}|\beta_j^{lasso} - \beta_j^{true}| \end{aligned} \tag{10}$$

and

$$\sum_{j\in\mathcal{P}} P'_\lambda(\beta_j^{true}) - \sum_{j\in\mathcal{P}} P'_\lambda(\beta_j^{lasso}) \leq \sum_{j\in\mathcal{P}} P'_\lambda(\beta_j^{lasso})(|\beta_j^{true}| - |\beta_j^{lasso}|) \leq \sum_{j\in\mathcal{P}} \lambda|\beta_j^{lasso} - \beta_j^{true}| \tag{11}$$

We also need 2 results from the proof for ϕ_0 in Corollary 2 which shows that both $|\delta^\ell_{\mathcal{S}_c}| \leq 3|\delta^\ell_{\mathcal{S}}|$ and $\frac{b_l}{n}\left\|X\delta^\ell\right\|^2 \leq 3\lambda^{lasso}|\delta^\ell_{\mathcal{S}}|$ conditional on $\mathcal{A}$ where $\delta^\ell = \beta^{lasso} - \beta^{true}$. Given our restricted eigenvalue assumption $\frac{\left\|X\delta^\ell\right\|^2}{n\left\|\delta^\ell\right\|^2} \geq r_e$, this can be used to show

$$\begin{aligned} |\delta^\ell| \leq 4|\delta^\ell_{\mathcal{S}}| &\leq 4\sqrt{s}\frac{\left\|\delta^\ell_{\mathcal{S}}\right\|^2}{\left\|\delta^\ell_{\mathcal{S}}\right\|} \leq 4\sqrt{s}\frac{\left\|\delta^\ell\right\|^2}{\left\|\delta^\ell_{\mathcal{S}}\right\|} \\ &\leq \frac{4\sqrt{s}}{r_e n}\frac{\left\|X\delta^\ell\right\|^2}{\left\|\delta^\ell_{\mathcal{S}}\right\|} \leq \frac{4\sqrt{s}}{r_e}\frac{3\lambda^{lasso}|\delta^\ell_{\mathcal{S}}|}{b_l\left\|\delta^\ell_{\mathcal{S}}\right\|} \leq \frac{4\sqrt{s}}{r_e}\frac{3\lambda^{lasso}\sqrt{s}\left\|\delta^\ell_{\mathcal{S}}\right\|}{b_l\left\|\delta^\ell_{\mathcal{S}}\right\|} \end{aligned} \tag{12}$$

which means $|\delta^\ell| \leq \frac{12\lambda^{lasso}s}{b_l r_e}$ with conditional on $\mathcal{A}$ which occurs with probability at least $1 - 2p\exp(\frac{-(\lambda^{lasso})^2 n b_u a}{8\sigma^2})$

Finally we are able to bound gamma by combining the above

$$\Gamma \leq \mathcal{Q}(\beta^*) - \mathcal{Q}(\beta^{true}) \leq \mathcal{Q}(\beta^{lasso}) - \mathcal{Q}(\beta^{true}) \tag{13}$$

$$\leq \sum_{i\in\mathcal{N}} \ell(\beta_j^{lasso}, x_i, y_i) - \sum_{j\in\mathcal{P}} P_\lambda(\beta_j^{lasso}) - [\sum_{i\in\mathcal{N}} \ell(\beta_j^{true}, x_i, y_i) - \sum_{j\in\mathcal{P}} P_\lambda(\beta_j^{true})] \tag{14}$$

$$\leq \sum_{j\in\mathcal{P}} (\lambda^{lasso}|\beta_j^{lasso} - \beta_j^{true}| + \lambda|\beta_j^{lasso} - \beta_j^{true}|) \tag{15}$$

$$\leq (\lambda^{lasso} + \lambda)|\delta^\ell| \leq (\lambda^{lasso} + \lambda)\frac{12\lambda^{lasso}s}{b_l r_e} \tag{16}$$

Next consider the conditions necessary to apply Theorem 1. We have assumptions (A1) and (A2) and (A3) per our assumption that the RE condition holds combined with 7. That leaves the 3 requirements on Γ, $P_\lambda(a\lambda)$ and $\|\beta_{\mathcal{S}}^{true}\|_{min}$. We will convert each of these to inequalities on n

Utilizing 16 and substituting $\lambda = \frac{Q\sigma}{r_e}\sqrt{\frac{\ln p}{n^{\gamma/2}}}$ and $\lambda^{lasso} = \epsilon\sigma\sqrt{\frac{\ln p}{n^{1-\gamma/2}}}$, where $Q, \epsilon > 0$ are arbitrary constants, and setting $\tilde{p}^* = 4s$, $t = \tilde{p}^* n^{\gamma/2}\ln p$, $t' = n^{\gamma/2}\ln p$ we get the following

$$P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1 + 2\sqrt{t'} + 2t') + \frac{\frac{\sigma^2}{2}s(1 + 2\sqrt{t'} + 2t') + \Gamma b_l}{b_l(\tilde{p}^* - 2s + 1)} \tag{17}$$

$$n > \frac{8 + 12\epsilon^2 + 12\epsilon Q}{b_l a Q^2} = C_1/b_l \tag{18}$$

$$\Gamma < P_\lambda(a\lambda) - \frac{\sigma^2}{b_l n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) \tag{19}$$

$$n > \left[\left(\frac{12\epsilon}{aQ} + \sqrt{\frac{20 + 12\epsilon^2}{aQ^2}}\right)\frac{s}{b_l}\right]^{\frac{2}{1-\gamma}} = \left[C_2\frac{s}{b_l}\right]^{\frac{2}{1-\gamma}} \tag{20}$$

$$\|\beta_{\mathcal{S}}^{true}\|_{\min} > \sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}^* + 2\sqrt{\tilde{p}^* t} + 2t\right) + \frac{8}{r_{\tilde{p}} b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathcal{S}|, P_\lambda(a\lambda)|\mathcal{S}| + \Gamma\}} \tag{21}$$

$$n > \left[(160 + 8Q^2)\frac{s\sigma^2 \ln p}{(\|\beta_{\mathcal{S}}^{true}\|_{min}\, r_{4s} b_l r_e)^2}\right]^{2/\gamma} = \left[C_3 \frac{s\sigma^2 \ln p}{(\|\beta_{\mathcal{S}}^{true}\|_{min}\, r_{4s} b_l r_e)^2}\right]^{2/\gamma} \tag{22}$$

For some constants C_1, C_2 and C_3

We can then apply Theorem 1 (conditional on $\mathcal{A}$) substitute our values and simplify to get that β^* is the global solution with probability at least

$$\begin{aligned}
&1 - 2\exp(-t + \tilde{p}^* \ln(\frac{pe}{\tilde{p}^*})) - 2\exp(-(\tilde{p}^* + 1)(t' - \ln p)) \cdot \left[\frac{1 - \exp(-(p - \tilde{p}^*)(t' - \ln p))}{1 - \exp(-t' + \ln p)}\right] \\
\geq& 1 - 2\exp(-(n^{\gamma/2} - 1)4s\ln p) - 2\left[\sum_{k=1}^{p-\tilde{p}^*} \exp(-(\tilde{p}^* + k)(n^{\gamma/2} - 1)\ln p)\right] \\
\geq& 1 - 2\exp(-(n^{\gamma/2} - 1)4s\ln p) - 2\exp(-[(4s+1)(n^{\gamma/2} - 1) - 1]\ln p) \\
\geq& 1 - C_4\exp(-C_5 s n^{\gamma/2}\ln p)
\end{aligned} \tag{23}$$

We then use the same technique as in Theorem 1 to combine this number with the probability of $\mathcal{A}$ to get the final non-conditional probability that β^* is the global solution with probability at least

$$\begin{aligned}
&\geq 1 - C_4\exp(-C_5 s n^{\gamma/2}\ln p) - 2p\exp(\frac{-(\lambda^{lasso})^2 n b_u a}{8\sigma^2}) \\
&\geq 1 - C_4\exp(-C_5 s n^{\gamma/2}\ln p) - 2\exp(\frac{-(\epsilon^2 b_u a n^{\gamma/2} - 8)\ln p}{8}) \\
&\geq 1 - C_4\exp(-C_5 s n^{\gamma/2}\ln p) - C_6\exp(-C_7 b_u n^{\gamma/2}\ln(p))
\end{aligned} \tag{24}$$

For some constants C_4, C_5 C_6 and C_7.

Note that these constants, as well as C_1, C_2 and C_3, are dependent only on the value of a Q and ϵ, as far as problem dependencies are concerned. Thus given that a Q and ϵ are chosen to be any positive constant value, as in the statement of Corollary 1, C_1 through C_7 are problem independent, which is the desired result. □

Proof of Corollary 2. The first result goes with the proof of Corollary 2 in Fan et al. (2014b). If we initialize the LLA algorithm with β^{lasso}, the solution to LASSO using λ^{lasso} as the LASSO constant, then the LLA algorithm converges to the oracle solution in 2 iterations with probability $1 - \phi_0 - \phi_1 - \phi_2$. The actual values of ϕ_0, ϕ_1, ϕ_2 are as follows.

First consider $\phi_0 = P(\left\|\beta^{lasso} - \beta^{true}\right\|_{max} > a_0\lambda)$. To bound this we will start by noticing that for the lasso penalized loss function $\mathcal{Q}^{lasso}(\beta) = \sum_{i\in\mathcal{N}} l(\beta, x_i, y_i) + \lambda^{lasso}\sum_{j\in\mathcal{P}}|\beta_j|$ we have that $\mathcal{Q}^{lasso}(\beta^{lasso}) \leq \mathcal{Q}^{lasso}(\beta^{true})$. If we then let $\delta^\ell = \beta^{lasso} - \beta^{true}$ we can use the same tactic as in the derivation of 8 to get $\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 - \frac{1}{n}W^\intercal X\delta^\ell \leq \lambda^{lasso}\sum_{j\in\mathcal{P}}|\beta_j^{true}| - |\beta_j^{lasso}|$, which can then be rearranged to get.

$$\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 - \frac{1}{n}\sum_{j\in\mathcal{P}}|W^\intercal X_j||\delta_j'| \leq \lambda^{lasso}\sum_{j\in\mathcal{P}}|\beta_j^{true}| - |\beta_j^{lasso}|. \tag{25}$$

Next let $\mathcal{A} = \bigcap_{j\in\mathcal{P}}\{|\frac{1}{n}W^\intercal X_j| \leq \lambda^{lasso}/2\}$. We can combine this with 25 to get that $\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 + \lambda^{lasso}/2\sum_{j\in\mathcal{P}}|\beta_j^{lasso} - \beta_j^{true}| \leq \lambda^{lasso}\sum_{j\in\mathcal{P}}|\beta_j^{lasso} - \beta_j^{true}| + \lambda^{lasso}\sum_{j\in\mathcal{P}}|\beta_j^{true}| - |\beta_j^{lasso}|$ conditional on $\mathcal{A}$. From this notice that the right term goes to zero when $\beta_j^{true} = 0$ so we then have that $\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 + \lambda^{lasso}/2\sum_{j\in\mathcal{P}}|\beta_j^{lasso} - \beta_j^{true}| \leq \lambda^{lasso}\sum_{j\in\mathcal{S}}|\beta_j^{lasso} - \beta_j^{true}| + |\beta_j^{true}| - |\beta_j^{lasso}|$. Using the triangle inequality and the definition of δ^ℓ we can simplify this to

$$\frac{b_l}{2n}\left\|X\delta^\ell\right\|^2 + \frac{\lambda^{lasso}}{2}|\delta^\ell| \leq 2\lambda^{lasso}|\delta^\ell_{\mathcal{S}}| \tag{26}$$

conditional on $\mathcal{A}$. By relaxing different parts of the equation, this can be further simplified to both $\frac{b_l}{n}\left\|X\delta^\ell\right\|^2 \leq 3\lambda^{lasso}|\delta^\ell_{\mathcal{S}}| \leq 3\lambda^{lasso}s^{1/2}||\delta^\ell_{\mathcal{S}}||_2$ and $|\delta^\ell_{\mathcal{S}_c}| \leq 3|\delta^\ell_{\mathcal{S}}|$. Note that the second of these shows that δ^ℓ satisfies the constraint for the RE condition 1. Therefor we have that $\frac{\left\|X\delta^\ell\right\|^2}{n\left\|\delta^\ell\right\|^2} \geq r_e$. If this is combined with the first of the two equations, we can get that $\frac{1}{n^{1/2}}\left\|X\delta^\ell\right\| \leq \frac{3\lambda^{lasso}s^{1/2}}{b_l(r_e)^{1/2}}$ conditional on $\mathcal{A}$.

Next, using this we can show that conditional on $\mathcal{A}$ we have that $\left\|\delta^\ell\right\|_{max} \leq \left\|\delta^\ell\right\| \leq \left\|X\delta^\ell\right\|_2^2/(\left\|\delta^\ell\right\| nr_e) \leq \frac{3\lambda^{lasso}s^{1/2}}{b_l r_e} < a_0\lambda$ if $\lambda > \frac{3\lambda^{lasso}s^{1/2}}{b_l a_0 r_e}$. This is the inverse of the condition that defines ϕ_0. Thus, we can bound ϕ_0 with $\phi_0 \leq P(\mathcal{A}^c) = P(\bigcup_{j\in\mathcal{P}}|\frac{1}{n}W^\intercal X_j| > \lambda^{lasso}/2) = P(\bigcup_{j\in\mathcal{P}}|W^\intercal X_j|/\left\|X_j\right\| > n\lambda^{lasso}/(2\left\|X_j\right\|)) \leq pP(|\langle W, v\rangle| > \frac{\lambda^{lasso}n}{2\|X_j\|}) \leq pP(|\langle W, v\rangle| > \frac{\lambda^{lasso}(nb_u a)^{1/2}}{2}) \leq 2p\exp\frac{-(\lambda^{lasso})^2 nb_u a}{8\sigma^2}$ which uses both (A1)(ii) and (A2) as long as $\lambda > \frac{3\lambda^{lasso}s^{1/2}}{b_l a_0 (r_e)}$ per (A2).

Next consider $\phi_1 = P(\left\|\nabla_{S^c_{\tilde{p}}}\ell_n(\beta^{oracle}\right\|_{max} \geq a_1\lambda)$

$$\begin{aligned}
\phi_1 =& P(\left\|\nabla_{S^c_{\tilde{p}}}\ell_n(\beta^{oracle}\right\|_{max} \geq a_1\lambda) && (27)\\
=& P(\exists j\in\mathcal{P} : |\nabla_j\ell_n(\beta^{oracle})| \geq a_1\lambda) && (28)\\
=& P(\exists j\in\mathcal{P} : |\frac{1}{n}\sum_{i\in\mathcal{N}}[\psi'(x_i^\intercal\beta^{oracle})x_{i,j} - y_i x_{i,j}]| \geq a_1\lambda) && (29)\\
=& P(\exists j\in\mathcal{P} : |\frac{1}{n}\sum_{i\in\mathcal{N}}[\psi'(x_i^\intercal\beta^{oracle})x_{i,j} - \psi'(x_i^\intercal\beta^{true})x_{i,j} + W_i x_{i,j}]| \geq a_1\lambda) && (30)\\
\leq& P(\frac{1}{n}|X_j^\intercal(\psi'(X\beta^{oracle}) - \psi'(X\beta^{true}) + W)| \geq a_1\lambda) && (31)\\
\leq& P(\frac{1}{n}|X_j^\intercal(\psi'(X\beta^{oracle}) - \psi'(X\beta^{true}))| + |W^\intercal X_j| \geq a_1\lambda) && (32)\\
\leq& P(\frac{1}{n}\left\|X_j\right\|\left\|\psi'(X\beta^{oracle}) - \psi'(X\beta^{true})\right\| + |W^\intercal X_j| \geq a_1\lambda) && (33)\\
\leq& P(\frac{1}{n}\left\|\psi'(X\beta^{oracle}) - \psi'(X\beta^{true})\right\| + |W^\intercal X_j|/\left\|X_j\right\| \geq a_1\lambda\left\|X_j\right\|^{-1}) && (34)\\
\leq& P(\left\|\psi'(X\beta^{oracle}) - \psi'(X\beta^{true})\right\| + |W^\intercal X_j|/\left\|X_j\right\| \geq (ab_u n)^{1/2}a_1\lambda) && (35)\\
\leq& P(b_u\left\|X\beta^{oracle} - X\beta^{true}\right\| + |W^\intercal v| \geq (ab_u n)^{1/2}a_1\lambda) && (36)\\
\leq& P(b_u\left\|X\delta^o\right\| + |W^\intercal v| \geq (ab_u n)^{1/2}a_1\lambda) && (37)
\end{aligned}$$

where $v \in \Re^n$ is some vector with $\|v\| = 1$ as indicated in (A2) and $\delta^o = \beta^{oracle} - \beta^{true}$.

From this, using Demorgan's law and the union bound, we notice that $P(A + B \geq C) \leq P(A \geq C/2) + P(B \geq C/2)$ which can be used to further simplify

$$\begin{aligned}
\phi_1 \leq& P(b_u\left\|X\delta^o\right\| + |W^\intercal v| \geq a_1\lambda(ab_u n)^{1/2}) && (38)\\
\leq& P(\|X\delta^o\| \geq (1/2)a_1\lambda(an/b_u)^{1/2}) + P(|W^\intercal v| \geq (1/2)a_1\lambda(ab_u n)^{1/2}) && (39)
\end{aligned}$$

We can then simplify both terms individually. For the first term, $P(b_u \|X\delta^o\| \geq (1/2)a_1\lambda(ab_u n)^{1/2})$, given the fact that the oracle solution and true solution have the same support, the oracle solution must be in the $\Gamma = 0$ level set of the true solution. Using similar arguments to Lemma 5, we have that $\frac{b_l}{2n}\|X\delta^o\|^2 \leq \frac{1}{n}W^\intercal X\delta^o$ From here Lemma 2 can be applied since we know $\left\|\beta^{oracle} - \beta^{true}\right\|_0 \leq s$. With some simplification this gives that $\|X\delta^o\| \leq \frac{2}{b_l}(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=s}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|)$ Utilizing Lemma 3 with s in place of $\tilde{p}$ shows that $P\left[\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=s}\frac{2}{b_l}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| \geq \frac{2}{b_l}\sigma\sqrt{s+2\sqrt{st_1}+2t_1}\right] \leq (\frac{pe}{s})^s\exp(-t_1)$. This is the first half of ϕ_1 as long as $(1/2)a_1\lambda(an/b_u)^{1/2} \geq \frac{2}{b_l}\sigma\sqrt{s+2\sqrt{st}+2t}$ which is equivalent to the assumed condition $\lambda \geq \frac{4\sigma\sqrt{s+2\sqrt{st_1}+2t_1}}{b_l a_1 (an/b_u)^{1/2}}$ Next, the second term can be easily bounded using (A2): $P(|W^\intercal v| \geq (1/2)a_1\lambda(ab_u n)^{1/2}) \leq 2\exp(-\frac{a_1^2\lambda^2 ab_u n}{8\sigma^2})$

Therefor $\phi_1 \leq (\frac{pe}{s})^s\exp(-t_1) + 2\exp(\frac{-a_1^2\lambda^2 ab_u n}{8\sigma^2})$

Next consider $\phi_2 = P(\left\|\beta_{\mathcal{S}}^{oracle}\right\|_{\min} \leq a\lambda)$ First, given the assumption $\left\|\beta^{true]}\right\|_{min} > (a+1)\lambda$ we can see that $\phi_2 = P(\left\|\beta_{\mathcal{S}}^{oracle}\right\|_{\min} \leq a\lambda) \leq P(\left\|\beta_{\mathcal{S}}^{oracle} - \beta_{\mathcal{S}}^{true}\right\|_{\max} > \lambda) \leq P(\left\|\beta^{oracle} - \beta^{true}\right\|_2 > \lambda) = P(\|\delta^o\|_2 > \lambda)$. Next, since we know that the support of β^{oracle} and β^{true} is $\mathcal{S}$, we know that $|\delta_{\mathcal{S}_c}^o| = 0 \leq 3|\delta_{\mathcal{S}}^o|$ which is the constraint for the RE condition. Therefor we know that $\frac{\|X\delta^o\|^2}{n\|\delta^o\|^2} \geq r_e$. With this and a similar line of argument as in ϕ_1 we get that $\phi_2 \leq P(\|\delta^o\| > \lambda) \leq P(\|X\delta^o\| > \lambda\sqrt{nr_e}) \leq P(\frac{2}{b_l}(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=s}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| > \lambda\sqrt{nr_e}) = P(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=s}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| > \lambda\frac{b_l\sqrt{nr_e}}{2} \geq \sigma\sqrt{s+2\sqrt{st_2}+2t_2}) \leq (\frac{pe}{s})^s\exp(-t_2)$ assuming that $\lambda\frac{b_l\sqrt{nr_e}}{2} \geq \sigma\sqrt{s+2\sqrt{st}+2t}$ which is equivalent to the condition $\lambda \geq \frac{2\sigma\sqrt{s+2\sqrt{st_2}+2t_2}}{b_l\sqrt{nr_e}}$

This, combined with the fact that for MCP, $a_0 = a_1 = a_2 = 1$ shows the first result.

The second result can be seen by first noting all assumptions of Theorem 1 part 2 are satisfied, where (A3) with r_{4s} is implied by 7. Thus by using the same arguments as in Theorem 1 part 2 which shows that the oracle solution is unique and that the global solution is the oracle solution with some probability, since the global solution is almost surely S^3ONC with $\Gamma = 0$. If we use $t = t_3$ and $t' = t_4$ we get that the probability that the global solution is not the oracle solution as $\phi_3 \leq \exp(-t_3 + \tilde{p}\ln(\frac{pe}{\tilde{p}})) + \exp(-(\tilde{p}^*+1)(t_4 - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p}^*)(t_4-\ln p))}{1-\exp(-t_4+\ln p)}$ This combined with the first result shows that the LLA algorithm converges to the global solution in 2 iterations with probability $1 - \phi_0 - \phi_1 - \phi_2 - \phi_3$ which is the second result. □

A.2 Central lemmas and their proofs

Lemma 1. *Let β^* be a S^3ONC solution to 2. If assumption (A1) holds, then $P[|\beta_j^*| \notin (0, a\lambda), \forall j \in \{1, 2, ..., p\}] = 1$.*

Proof of Lemma 1. First, define events γ_j and δ_j as

$$\gamma_j := \left\{ \left.\frac{\partial^2\mathcal{Q}(\beta)}{(\partial\beta_j)^2}\right|_{\beta=\beta^*} \geq 0 \right\} \tag{40}$$

$$\delta_j := \left\{|\beta_j^*| \in (0, a\lambda)\right\}. \tag{41}$$

First, for any given $j \in \mathcal{P}$, we solve for $P[\gamma_j \cap \delta_j]$ given our assumptions. We can start with $\left.\frac{\partial^2\mathcal{Q}(\beta)}{(\partial\beta_j)^2}\right|_{\beta=\beta^*} \geq 0$ which gives us $1/n\sum_{i=1}^n \psi''\left(x_i^\intercal\beta^*\right)x_{i,j}^2 + P_\lambda''(|\beta_j^*|) \geq 0$. We can rearrange this to get $b_u\sum_{i=1}^n x_{i,j}^2 \geq \sum_{i=1}^n \psi''(x_i^\intercal\beta^*)x_{i,j}^2 \geq -nP_\lambda''(|\beta^*|) = n/a$ where we get the leftmost inequality from assumption (A1) part (i) and the rightmost equality from the definition of MCP. More concisely we have that $b_u \|X_j\|^2 \geq n/a$ which contradicts (A1) part (ii). Therefor we know $P[\gamma_j \cap \delta_j] = 0$. It should also be noted that $P[\gamma_j^c] = 0$ since β^* satisfies S^3ONC conditions. Thus, by applying Demorgan's law and then the union bound, it can be obtained that

$$0 = P[\gamma_j \cap \delta_j] = 1 - P[\gamma_j^c \cup \delta_j^c] \geq 1 - P[\gamma_j^c] - P[\delta_j^c] = 1 - P[\delta_j^c] = P[\delta_j]. \tag{42}$$

We can then apply this result to all indices to get that $P[\delta_j] = 0$ for all $j \in \{1, 2, ..., p\}$, which is the desired result. □

Lemma 2. *Consider an arbitrary S^3ONC solution β^* to 2 with MCP. Given the event that for some integer $\tilde{p}$: $\|\beta^* - \beta^{true}\|_0 \leq \tilde{p}$, then $|W^\intercal X\delta^*| \leq \left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right) \|X\delta^*\|, \quad a.s.$*

Where

$$(\tilde{U}_{S_{\tilde{p}}})_{i,j} := \begin{cases} U_{S_{\tilde{p}}}, & \text{if } j \in S_{\tilde{p}} \\ 0, & \text{else} \end{cases}$$

and $U_{S_{\tilde{p}}} \in \Re^{n\times\tilde{p}}$ is defined as in the following Thin SVD: $X_{S_{\tilde{p}}} = U_{S_{\tilde{p}}} D_{S_{\tilde{p}}} V_{S_{\tilde{p}}}$.

Proof. Denote $\delta^* := (\delta_j^*) = \beta^* - \beta^{true}$, $S_{\tilde{p}} := (j : \delta_j^* \neq 0) \subseteq \mathcal{P}$, $\delta_{S_{\tilde{p}}}^* := (\delta_j^* : j \in S_{\tilde{p}})$ and $X_{S_{\tilde{p}}} := (x_{ij} : i \in \mathcal{N}, j \in S_{\tilde{p}})$. By assumption, we know that $\|\delta^*\|_0 \leq |S_{\tilde{p}}| = \tilde{p}$.

First decompose $X_{S_{\tilde{p}}}$ using Thin SVD to get $X_{S_{\tilde{p}}} = U_{S_{\tilde{p}}} D_{S_{\tilde{p}}} V_{S_{\tilde{p}}}$ where $U_{S_{\tilde{p}}} \in \Re^{n\times\tilde{p}}$. Note that since and $U_{S_{\tilde{p}}}^\intercal U_{S_{\tilde{p}}} = I$ we have that for any $\upsilon \in \Re^{\tilde{p}}$ we have $\left\|D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \upsilon\right\|^2 = (D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \upsilon)^\intercal I (D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \upsilon) = \upsilon^\intercal V_{S_{\tilde{p}}}^\intercal D_{S_{\tilde{p}}}^\intercal U_{S_{\tilde{p}}}^\intercal U_{S_{\tilde{p}}} D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \upsilon = \upsilon^\intercal X_{S_{\tilde{p}}}^\intercal X_{S_{\tilde{p}}} \upsilon = \left\|X_{S_{\tilde{p}}} \upsilon\right\|^2$. therefor we can obtain that

$$\begin{aligned} |W^\intercal X\delta^*| &= |W^\intercal X_{S_{\tilde{p}}} \delta_{S_{\tilde{p}}}^*| \leq \left\|W^\intercal U_{S_{\tilde{p}}}\right\| \left\|D_{S_{\tilde{p}}} V_{S_{\tilde{p}}} \delta_{S_{\tilde{p}}}^*\right\| \\ &= \left\|U_{S_{\tilde{p}}}^\intercal W\right\| \left\|X_{S_{\tilde{p}}} \delta_{S_{\tilde{p}}}^*\right\| \leq \left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right) \|X\delta^*\|, \quad a.s. \end{aligned} \tag{43}$$

Where

$$(\tilde{U}_{S_{\tilde{p}}})_{i,j} := \begin{cases} U_{S_{\tilde{p}}}, & \text{if } j \in S_{\tilde{p}} \\ 0, & \text{else} \end{cases}.$$

□

Lemma 3. *Consider an arbitrary S^3ONC solution β^* to 2 with MCP. If (A2) holds, then for some integer $\tilde{p} \leq p$, $P\left[\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| \leq \sigma\sqrt{\tilde{p} + 2\sqrt{\tilde{p}t} + 2t}\right] \geq 1 - (\frac{pe}{\tilde{p}})^{\tilde{p}} \exp(-t)$. Where*

$$(\tilde{U}_{S_{\tilde{p}}})_{i,j} := \begin{cases} U_{S_{\tilde{p}}}, & \text{if } j \in S_{\tilde{p}} \\ 0, & \text{else} \end{cases}$$

and $U_{S_{\tilde{p}}} \in \Re^{n\times\tilde{p}}$ is defined as in the following Thin SVD: $X_{S_{\tilde{p}}} = U_{S_{\tilde{p}}} D_{S_{\tilde{p}}} V_{S_{\tilde{p}}}$.

Proof. We attempt to bound $\left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right)$. Given that we now have W multiplied by a square matrix, we can apply Lemma 9. In the Lemma, let $\Sigma_u = \tilde{U}_{S_{\tilde{p}}} \tilde{U}_{S_{\tilde{p}}}^\intercal$. The fact that $\Sigma_u \Sigma_u = \Sigma_u$ means that Σ_u is an idempotent matrix with $\|\Sigma_u\| \leq 1$ and $Tr(\Sigma_u) = rank(\Sigma_u) \leq rank(\tilde{U}_{S_{\tilde{p}}}) \leq rank(U_{S_{\tilde{p}}}) \leq \tilde{p}$. Lemma 9 then states that that $P\left[\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| \leq \sigma\sqrt{\tilde{p} + 2\sqrt{\tilde{p}t} + 2t}\right] \geq 1 - \exp(-t)$. From this we can show that

$$P\left[\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}} \left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| \leq \sigma\sqrt{\tilde{p} + 2\sqrt{\tilde{p}t} + 2t}\right] \geq 1 - \binom{p}{\tilde{p}} \exp(-t) \geq 1 - (\frac{pe}{\tilde{p}})^{\tilde{p}} \exp(-t). \tag{44}$$

Where the first inequality can seen by noting that if $\eta_k \in \Re^k$ is a sequence of i.i.d random variables and $\theta \in \Re$ is a scalar, by applying De Morgan's Law and then using the union bound, it can be obtained that $P[\max_{k\in K} \eta_k \leq \theta] = P[\bigcap_{k\in K} \eta_k \leq \theta] = 1 - P[\bigcup_{k\in K} \eta_s \geq \theta] \geq 1 - \sum_{k\in K} P[\eta_k \geq \theta] = 1 - |K|(1 - P[\eta_k \leq \theta])$ which yields the same inequality as in 44.

This is the desired result. □

Lemma 4. *Consider an arbitrary S^3ONC solution β^* to 2 with MCP. Let Assumptions (A1) and (A2) hold. Given the simultaneous occurrence of (i) the event that $\mathcal{Q}(\beta^*) \leq \mathcal{Q}(\beta^{true}) + \Gamma$ holds for some $\Gamma \geq 0$; (ii) the event that for some integer $\tilde{p} : \|\beta^* - \beta^{true}\|_0 \leq \tilde{p}$. Then for any $t > 0$, $\frac{1}{n}\|X(\beta^* - \beta^{true})\|^2 \leq \frac{4\sigma^2}{b_l^2 n}(\tilde{p} + 2\sqrt{\tilde{p}t} + 2t) + \frac{8}{b_l}\min\{\sum_{j\in S} P'_\lambda(|\beta^*_j|)|\beta^*_j|, P_\lambda(a\lambda)(|\mathcal{S}| - \|\beta^*\|_0) + \Gamma\}$ holds with probability at least $1 - \exp(-t + \tilde{p}\ln(\frac{pe}{\tilde{p}}))$.*

If in addition (A3) holds with $\tilde{p}^* \geq \tilde{p}$, then $\frac{1}{n}\|X(\beta^* - \beta^{true})\|^2 \leq \frac{8\sigma^2}{b_l^2 n}\left(\tilde{p} + 2\sqrt{\tilde{p}t} + 2t\right) + \frac{8}{b_l}\min\{\lambda^2(|\mathcal{S}| - \|\beta^*\|_0)r_{\tilde{p}}^{-1}, P_\lambda(a\lambda)\dot{(}|\mathcal{S}| - \|\beta^*\|_0) + \Gamma\}$ holds where $r_{\tilde{p}} > 0$ for any $t > 0$ with probability at least $1 - \exp(-t + \tilde{p}\ln(\frac{pe}{\tilde{p}}))$.

Proof. First, denote $\delta^* := (\delta^*_j) = \beta^* - \beta^{true}$, $S_{\tilde{p}} := (j : \delta^*_j \neq 0) \subseteq \mathcal{P}$, $\delta^*_{S_{\tilde{p}}} := (\delta^*_j : j \in S_{\tilde{p}})$ and $X_{S_{\tilde{p}}} := (x_{ij} : i \in \mathcal{N}, j \in S_{\tilde{p}})$. By assumption, we know that $\|\delta^*\|_0 \leq |S_{\tilde{p}}| = \tilde{p}$. Further, let us denote

$$\mathcal{T}_1 := \min\left\{\sum_{j\in S} P'_\lambda(|\beta^*_j|)|\beta^{true}_j|,\ \sum_{j\in S} P'_\lambda(|\beta^*_j|)|\beta^*_j - \beta^{true}_j|,\ P_\lambda(a\lambda)(|\mathcal{S}| - \|\beta^*\|_0) + \Gamma\right\}. \quad (45)$$

We now start to define the desired bound by applying the second part of Lemma 8. The result simplified using the above definitions becomes

$$\frac{b_l}{2n}\|X\delta^*\|^2 \leq \frac{1}{n}W^\intercal X\delta^* + \mathcal{T}_1, \quad a.s. \quad (46)$$

Next, since all assumptions for Lemma 1 are satisfied, we can apply it to get

$$\frac{b_l}{2n}\|X\delta^*\|^2 \leq \frac{1}{n}\left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}^\intercal_{S_{\tilde{p}}}W\right\|\right)\|X\delta^*\| + \mathcal{T}_1. \quad (47)$$

We can then complete the square, solving for $\frac{1}{\sqrt{n}}\|X\delta^*\|$ to get

$$\frac{1}{\sqrt{n}}\|X\delta^*\| \leq \frac{1}{b_l\sqrt{n}}\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}^\intercal_{S_{\tilde{p}}}W\right\| + \sqrt{\left(\frac{1}{b_l\sqrt{n}}\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}^\intercal_{S_{\tilde{p}}}W\right\|\right)^2 + \frac{2}{b_l}\mathcal{T}_1} \quad (48)$$

$$\leq 2\sqrt{\left(\frac{1}{b_l\sqrt{n}}\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}^\intercal_{S_{\tilde{p}}}W\right\|\right)^2 + \frac{2}{b_l}\mathcal{T}_1}, \quad (49)$$

where the last inequality holds due to the value inside the square root being larger than the term outside. From here, squaring both sides gives us

$$\frac{1}{n}\|X\delta^*\|^2 \leq \frac{4}{b_l^2 n}\left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}^\intercal_{S_{\tilde{p}}}W\right\|\right)^2 + \frac{8}{b_l}\mathcal{T}_1. \quad (50)$$

Finally, by applying the second part of Lemma 1 we get

$$\frac{1}{n}\|X\delta^*\|^2 \leq \frac{4\sigma^2}{b_l^2 n}\left(\tilde{p} + 2\sqrt{\tilde{p}t} + 2t\right) + \frac{8}{b_l}\mathcal{T}_1. \quad (51)$$

With probability at least $1 - (\frac{pe}{\tilde{p}})^{\tilde{p}} \exp(-t)$. Thus by the definition of $\mathcal{T}_1$, the first result of the lemma has been shown.

For the second part we look to bound the central term of $\mathcal{T}_1$. We first notice (a) that since assumption (A1) holds, Corollary 4 indicates that if $\beta_j^* \neq 0 \Rightarrow |\beta_j^*| \geq a\lambda$ for all $j \in \mathcal{P}$; (b) that for this range of β_j^*, $P'_\lambda(|\beta_j^*|) = 0$; (c) that per the definition of MCP $0 \leq P'_\lambda(|\beta_j^*|) \leq \lambda$ for any $\beta_j^* \in \Re$. If we combine these observations with 45 and the definition of δ^*, we can see that $\mathcal{T}_1 \leq \sum_{j \in \mathcal{S}} P'_\lambda(|\beta_j^*|)|\delta^*| \leq \lambda\sqrt{|\mathcal{S}| - \|\beta_{\mathcal{S}}^*\|_0} \cdot \|\delta^*\|$. From this, given that assumption that, for this second result (A3) holds with $\tilde{p}^* \geq \tilde{p}$, and $r_{\tilde{p}} \geq r_{\tilde{p}^*} \geq 0$ we can use (A3) part (iii) to show that $\mathcal{T}_1 \leq \lambda\sqrt{|\mathcal{S}| - \|\beta_{\mathcal{S}}^*\|_0} \cdot \frac{\|X\delta^*\|}{\sqrt{nr_{\tilde{p}}}}$. Since this holds almost surely, it can then be combined with 47 to get

$$\frac{b_l}{2n}\|X\delta^*\|^2 \leq \frac{1}{n}\left(\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|\right)\|X\delta^*\| + \lambda\sqrt{|\mathcal{S}| - \|x_{\mathcal{S}}^*\|_0} \cdot \frac{\|X\delta^*\|}{\sqrt{nr_{\tilde{p}}}}. \tag{52}$$

We can then multiply by $2\sqrt{n}/b_l\,\|X\delta^*\|$ to get

$$\frac{1}{\sqrt{n}}\|X\delta^*\| \leq \frac{2}{b_l\sqrt{n}}\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| + \frac{2\lambda}{b_l\sqrt{r_{\tilde{p}}}}\sqrt{|\mathcal{S}| - \|x_{\mathcal{S}}^*\|_0}. \tag{53}$$

We then square both sides and use the rule that $(A+B)^2 \leq 2A^2 + 2B^2$ to get

$$\frac{1}{n}\|X\delta^*\|^2 \leq \left[\frac{2}{b_l\sqrt{n}}\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\| + \frac{2\lambda}{b_l\sqrt{r_{\tilde{p}}}}\sqrt{|\mathcal{S}| - \|x_{\mathcal{S}}^*\|_0}\right]^2 \tag{54}$$

$$\leq \frac{8}{b_l^2 n}\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|^2 + \frac{8\lambda^2}{b_l^2 r_{\tilde{p}}}(|\mathcal{S}| - \|x_{\mathcal{S}}^*\|_0). \tag{55}$$

Combining this with 50 yields that

$$\frac{1}{n}\|X\delta^*\|^2 \leq \frac{8}{b_l^2 n}\max_{S_{\tilde{p}}:|S_{\tilde{p}}|=\tilde{p}}\left\|\tilde{U}_{S_{\tilde{p}}}^\intercal W\right\|^2 + \frac{8}{b_l}\min\left\{\frac{\lambda^2}{r_{\tilde{p}}}(|\mathcal{S}| - \|x_{\mathcal{S}}^*\|_0), \mathcal{T}_1\right\}. \tag{56}$$

Finally, by applying the second part of Lemma 1 and noting from (45) that $\mathcal{T}_1 \leq P_\lambda(a\lambda)(|\mathcal{S}| - \|\beta^*\|_0) + \Gamma$ we see that

$$\frac{1}{n}\|X\delta^*\|^2 \leq \frac{8}{b_l^2 n}\sigma^2(\tilde{p} + 2\sqrt{\tilde{p}t} + 2t) + \frac{8}{b_l}\min\left\{\frac{\lambda^2}{r_{\tilde{p}}}(|\mathcal{S}| - \|x_{\mathcal{S}}^*\|_0), P_\lambda(a\lambda)(|\mathcal{S}| - \|\beta^*\|_0) + \Gamma\right\}. \tag{57}$$

With probability at least $1 - (\frac{pe}{\tilde{p}})^{\tilde{p}} \exp(-t)$ which is the desired result. □

Lemma 5. *Let Assumptions (A1) and (A2) hold. Consider a solution β^* satisfying S^3ONC of 2. Assume that $\mathcal{Q}(\beta^*) \leq \mathcal{Q}(\beta^{true}) + \Gamma$ holds for an arbitrary $\Gamma > 0$. For any integer $\tilde{p} : 2|\mathcal{S}| \leq \tilde{p} \leq p$ if the penalty parameters (a, λ) satisfy that $P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1 + 2\sqrt{t} + 2t) + \frac{\frac{\sigma^2}{n}|\mathcal{S}|(1+2\sqrt{t}+2t)+\Gamma b_l}{b_l(\tilde{p}+1-2|\mathcal{S}|)}$, for an arbitrary $t > 0$, then $\|\beta^* - \beta^{true}\|_0 \leq \tilde{p}$ with probability at least $1 - \exp(-(\tilde{p}+1)(t - \ln p)) \cdot \frac{1-\exp(-(p-\tilde{p})(t-\ln p))}{1-\exp(-t+\ln p)}$.*

Proof. We start from the useful inequality defined in 8

$$\frac{b_l}{2n}\|X\delta^*\|^2 - \frac{1}{n}W^\intercal X\delta^* + \sum_{j \in \mathcal{S}} P_\lambda(|\beta_j^*|) \leq \sum_{j \in \mathcal{S}} P_\lambda(|\beta_j^{true}|) + \Gamma, \tag{58}$$

where $\delta^* = \beta^* - \beta^{true}$. Next, conditioning on the fact (i) that β^* is S^3ONC, (ii) that all assumptions for Corollary 4 are satisfied (which implies that $P_\lambda(|\beta_j^*|) \in \{0, P_\lambda(a\lambda)\}$) and (iii) that $P_\lambda(|\beta_j^{true}|) \leq P_\lambda(a\lambda)$ we have that

$$\frac{b_l}{2n}\|X\delta^*\|^2 - \frac{1}{n}W^\intercal X\delta^* + \|\beta^*\|_0 \cdot P_\lambda(a\lambda) \leq |\mathcal{S}| \cdot P_\lambda(a\lambda) + \Gamma \tag{59}$$

Now consider an event $\mathcal{E}_1 := \{\|\beta^* - \beta^{true}\|_0 = \tilde{p} + k\}$ for an arbitrary integer $k : 1 \leq k \leq p - \tilde{p}$ Conditioning on this event, we may denote and $S_{\tilde{p}+k} \subseteq \mathcal{P}$ such that $\delta_j^* \neq 0$ for all $j \in S_{\tilde{p}+k}$. By assumption we can ensure that $|S_{\tilde{p}+k}| = \tilde{p} + k$. Also denote by $X_{S_{\tilde{p}+k}} = (x_{ij} : i \in \mathcal{N}, j \in S_{\tilde{p}+k})$ and let $\delta^*_{S_{\tilde{p}+k}} := (\delta_j^* : j \in S_{\tilde{p}+k})$. Note that conditional on $\mathcal{E}_1$, the first part of Lemma 1 (using $\tilde{p} + k$ in place of $\tilde{p}$ in Lemma 1) can be used to bound $W^\intercal X\delta^*$ in 59. Additionally, by definition $\|\beta^{true}\|_0 = |\mathcal{S}|$ and conditional on $\mathcal{E}_1$ we can apply the substitution $\|\beta^*\|_0 \geq \tilde{p} + k - |\mathcal{S}|$. This gives us

$$\frac{b_l}{2}\left\|\frac{X\delta^*}{\sqrt{n}}\right\|^2 - \frac{1}{\sqrt{n}}\left(\max_{S_{\tilde{p}+k}:|S_{\tilde{p}+k}|=\tilde{p}+k}\left\|\tilde{U}^\intercal_{S_{\tilde{p}+k}}W\right\|\right)\left\|\frac{X\delta^*}{\sqrt{n}}\right\| \leq -(\tilde{p}+k-2|\mathcal{S}|)\cdot P_\lambda(a\lambda)+\Gamma \tag{60}$$

In order for this equation to be feasible, we know that the quadratic formula must have real roots. Therefor

$$\left(\max_{S_{\tilde{p}+k}:|S_{\tilde{p}+k}|=\tilde{p}+k}\left\|\frac{\tilde{U}^\intercal_{S_{\tilde{p}+k}}W}{\sqrt{n}}\right\|\right)^2 - 4[b_l/2][(\tilde{p}+k-2|\mathcal{S}|)\cdot P_\lambda(a\lambda) - \Gamma] \geq 0 \tag{61}$$

Now consier another event $\mathcal{E}_2(t) := \{\max_{|S_{\tilde{p}+k}|=\tilde{p}+k} ||U^\intercal_{S_{\tilde{p}+k}}W|| \leq \sigma\sqrt{\tilde{p}+k}\cdot\sqrt{1+2\sqrt{t}+2t}\}$ for an arbitrary $t > 0$. Conditioning on the occurence of $\mathcal{E}_1 \cap \mathcal{E}_2(t)$ we can show, using first $\mathcal{E}_2(t)$ and then 61, that $\frac{\sigma^2(\tilde{p}+k)}{n}\cdot(1+2\sqrt{t}+2t) \geq \left(\max_{|S_{\tilde{p}+k}|=\tilde{p}+k}\frac{\left\|\tilde{U}^\intercal_{S_{\tilde{p}+k}}W\right\|}{\sqrt{n}}\right)^2 \geq 2b_l[(\tilde{p}+k-2|\mathcal{S}|)\cdot P_\lambda(a\lambda) - \Gamma]$ almost surely, which contradicts with the assumption on the parameters (a, λ). This can be seen starting from our original assumption that $P_\lambda(a\lambda) > \frac{\sigma^2}{2nb_l}(1+2\sqrt{t}+2t) + \frac{\frac{\sigma^2}{n}|\mathcal{S}|(1+2\sqrt{t}+2t)+\Gamma b_l}{b_l(\tilde{p}+1-2|\mathcal{S}|)} \geq \frac{\sigma^2}{2nb_l}(1+2\sqrt{t}+2t) + \frac{\frac{\sigma^2}{n}|\mathcal{S}|(1+2\sqrt{t}+2t)+\Gamma b_l}{b_l(\tilde{p}+k-2|\mathcal{S}|)}$ We can then multiply both (outer) sides by $2b_l(\tilde{p}+k-2|\mathcal{S}|)$ and rearrange to get $\frac{\sigma^2}{n}(\tilde{p}+k)\cdot(1+2\sqrt{t}+2t) < 2b_l[(\tilde{p}-2|\mathcal{S}|+k)\cdot P_\lambda(a\lambda) - \Gamma]$. Given this contradiction, we know that $P[\mathcal{E}_1 \cap \mathcal{E}_2(t)] = 0$. Therefore, again using the union bound combined with DeMorgan's law we get that $P[\mathcal{E}_1 \cap \mathcal{E}_2(t)] \geq 1 - P[\mathcal{E}_1^c] - P[\mathcal{E}_2(t)^c]$ move to intro somehwere? which, can be simplified to

$$P[\mathcal{E}_2(t)^c] \geq P[\mathcal{E}_1] \tag{62}$$

Since all assumptions of the Lemma 3 are satisfied, we can next use it to bound $P[\mathcal{E}_2(t)^c]$. By taking the compliment of the result in the second half of Lemma 1, we get, for some t', that $P\left[\max_{S_{\tilde{p}+k}:|S_{\tilde{p}+k}|=(\tilde{p}+k)}\left\|\tilde{U}^\intercal_{S_{\tilde{p}+k}}W\right\| \geq \sigma\sqrt{(\tilde{p}+k)+2\sqrt{(\tilde{p}+k)t'}+2t'}\right] \leq (\frac{pe}{\tilde{p}+k})^{\tilde{p}+k}\exp(-t') \leq p^{\tilde{p}+k}\exp(-t')$. If we then let $t' = (\tilde{p}+k)t$ we get $P\left[\max_{S_{\tilde{p}+k}:|S_{\tilde{p}+k}|=(\tilde{p}+k)}\left\|\tilde{U}^\intercal_{S_{\tilde{p}+k}}W\right\| \geq \sigma\sqrt{\tilde{p}+k}\cdot\sqrt{1+2\sqrt{t}+2t}\right] \leq p^{\tilde{p}+k}\exp(-(\tilde{p}+k)t)$. Thus we have that $p^{\tilde{p}+k}\exp(-(\tilde{p}+k)t) \geq P[\mathcal{E}_2(t)^c]$ which can be combined with 62 to show

$$p^{\tilde{p}+k}\exp(-(\tilde{p}+k)t \geq P[\|\beta^* - \beta^{true}\|_0 = \tilde{p}+k] \quad \forall k \in \mathbb{Z} : 1 \leq k \leq p - \tilde{p}. \tag{63}$$

With this, we can solve for our desired value

$$
\begin{aligned}
P\left[\left\|\beta^*-\beta^{true}\right\|_0 \le \tilde{p}\right] &= 1-P\left[\left\|\beta^*-\beta^{true}\right\|_0 \ge \tilde{p}+1\right] = 1-\sum_{k=1}^{p-\tilde{p}} P\left[\left\|\beta^*-\beta^{true}\right\|_0 = \tilde{p}+k\right] \\
&\ge 1-\sum_{k=1}^{p-\tilde{p}} \exp((\tilde{p}+k)(\ln p - t)) \\
&= 1-\exp(-(\tilde{p}+1)(t-\ln p))\cdot \frac{1-\exp(-(p-\tilde{p})(t-\ln p))}{1-\exp(-t+\ln p)}
\end{aligned}
\tag{64}
$$

Which is the desired result.

□

Lemma 6. *Consider an arbitrary S^3ONC solution β^* to 2 with MCP. Let Assumptions (A1) and (A3) with $\tilde{p}^* \ge \tilde{p}$ hold. Assume the satisfaction of $\|\beta^*-\beta^{true}\| \le \tilde{p}$ and Event $\mathcal{E}_a(\tilde{p}) := \{\frac{1}{n}\left\|X(\beta^*-\beta^{true})\right\|^2 \le \frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}(|\mathcal{S}|-\|\beta^*\|_0), P_\lambda(a\lambda)\cdot(|\mathcal{S}|-\|\beta^*\|_0)+\Gamma\}$. If the sub-optimality gap satisfies $\Gamma < P_\lambda(a\lambda)-\frac{\sigma^2}{b_l n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)$. If the minimum signal strength satisfies $\|\beta_{\mathcal{S}}^{true}\|_{\min} > \sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{r_{\tilde{p}} b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathcal{S}|, P_\lambda(a\lambda)|\mathcal{S}|+\Gamma\}}$ then β^* is the oracle solution to 2.*

If in addition we have the satisfaction of $\|\beta^{opt}-\beta^{true}\| \le \tilde{p}$ and the event $\mathcal{E}_b(\tilde{p}) := \{\frac{1}{n}\left\|X(\beta^{opt}-\beta^{true})\right\|^2 \le \frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}(|\mathcal{S}|-\|\beta^*\|_0), P_\lambda(a\lambda)\cdot(|\mathcal{S}|-\|\beta^*\|_0)+\Gamma\}$ then β^* is both the oracle solution and the global solution to 2.

Proof. First, let us denote $\beta^*-\beta^{true}=\delta^*$. We start by combining $\mathcal{E}_\alpha(\tilde{p})$ and (A3)iii, which is possible due to our assumption $\|\beta^*-\beta^{true}\|_0 \le \tilde{p}$. This gives us

$$
\begin{aligned}
\frac{8\sigma^2}{b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}(|\mathcal{S}|-\|\beta^*\|_0), P_\lambda(a\lambda)\cdot(|\mathcal{S}|-\|\beta^*\|_0)+\Gamma\} \\
\ge \frac{1}{n}\left\|X\delta^*\right\|^2 \ge r_{\tilde{p}}\left\|\delta^*\right\|^2 \;\; a.s.
\end{aligned}
\tag{65}
$$

From here if we relax $|\mathcal{S}|-\|\beta_{\mathcal{S}}^*\|_0$ to just $|\mathcal{S}|$, the definition of δ^* and note that $\|\delta^*\| \ge \left\|\delta_j^*\right\|$, we can obtain the following

$$
\begin{aligned}
\sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{r_{\tilde{p}} b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathcal{S}|, P_\lambda(a\lambda)|\mathcal{S}|+\Gamma\}} \\
\ge \left\|\beta_j^*-\beta_j^{true}\right\| \ge |\beta_j^{true}|-|\beta_j^*|,
\end{aligned}
\tag{66}
$$

almost surely. From this we can bound $|\beta_j^*|$ using the square root term and $|\beta_j^{true}|$, so we know that if $|\beta_j^{true}|-\sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{r_{\tilde{p}} b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathcal{S}|, P_\lambda(a\lambda)|\mathcal{S}|+\Gamma\}} > 0$ then $|\beta_j^*| > 0$. From this we can obtain the inequality

$$
\|\beta_{\mathcal{S}}^*\|_0 \ge \sum_{j\in\mathcal{S}} \mathbb{I}\left(|\beta_j^{true}|-\sqrt{\frac{8\sigma^2}{r_{\tilde{p}} b_l^2 n}\left(\tilde{p}+2\sqrt{\tilde{p}t}+2t\right)+\frac{8}{r_{\tilde{p}} b_l}\min\{\frac{\lambda^2}{r_{\tilde{p}}}|\mathcal{S}|, P_\lambda(a\lambda)|\mathcal{S}|+\Gamma\}} > 0\right)
\tag{67}
$$

almost surely. We can then combine this with our minimum signal strength assumption to get that

$$
\|\beta_{\mathcal{S}}^*\|_0 = |\mathcal{S}| \; a.s.
\tag{68}
$$

We can combine this with equation 65, by focusing on the second part of the minimum term and noting the right side is always positive to get

$$\frac{8\sigma^2}{b_l^2 n}\left(\tilde{p} + 2\sqrt{\tilde{p}t} + 2t\right) + \frac{8}{b_l}(-P_\lambda(a\lambda)\left\|\beta^*_{\mathcal{S}^c}\right\|_0 + \Gamma) \geq 0 \; a.s. \tag{69}$$

which can be simplified into

$$\frac{\sigma^2}{b_l n}\left(\tilde{p} + 2\sqrt{\tilde{p}t} + 2t\right) + \Gamma \geq P_\lambda(a\lambda)\left\|\beta^*_{\mathcal{S}^c}\right\|_0 \; a.s. \tag{70}$$

thus, it can be seen that if $P_\lambda(a\lambda) > \frac{\sigma^2}{b_l n}\left(\tilde{p} + 2\sqrt{\tilde{p}t} + 2t\right) + \Gamma$ then $1 > \left\|\beta^*_{\mathcal{S}^c}\right\|_0 = 0$. This is satisfied by the assumption that $P_\lambda(a\lambda) - \frac{\sigma^2}{b_l n}\left(\tilde{p} + 2\sqrt{\tilde{p}t} + 2t\right) > \Gamma$.

Finally, because β^* is an S^3ONC solution, it has to satisfy FONC. Per 2, this means that $\beta^* \in \arg\inf\{\frac{1}{n}\sum_{i\in\mathcal{N}}\ell(\beta, x_i, y_i) + \sum_{j\in\mathcal{P}} P'_\lambda(|\beta^*_j|)|\beta_j| : \; \beta \in \Re^p\}$. Due to Corollary 4, we know that the penalty term goes to 0 since either $\beta^*_j = 0$ or $P'(|\beta^*_j|) = P'(|a\lambda|) = 0$. Further we know that $\beta^*_j = 0$ for all $j \in \mathcal{S}^c$. Threfore we know that

$$\beta^* \in \arg\inf\{\frac{1}{n}\sum_{i\in\mathcal{N}}\ell(\beta, x_i, y_i) : \; \beta \in \Re^p, \beta_j = 0, \forall j \in \mathcal{S}^c\} \; a.s. \tag{71}$$

Given that the expression on the right is the definition of the oracle solution, we have shown the first result.

Next, Consider β^{opt} which is the global optimal solution to 2. Given that the S^3ONC conditions are necessary, β^{opt} must be an S^3ONC solution. With this fact and the assumption of $\mathcal{E}_b(\tilde{p})$, we have the same set of assumptions for β^{opt} as we had for β^*. Thus the same sequence of arguments can be used to show that

$$\beta^{opt} \in \arg\inf\{\frac{1}{n}\sum_{i\in\mathcal{N}}\ell(\beta, x_i, y_i) : \; \beta \in \Re^p, \beta_j = 0, \forall j \in \mathcal{S}^c\} \; a.s. \tag{72}$$

Finally, per the strict convexity of our loss function as implied by (A1) we can see that the infimum of the above problem is unique. Therefor

$$\beta^* = \arg\inf\{\frac{1}{n}\sum_{i\in\mathcal{N}}\ell(\beta, x_i, y_i) : \; \beta \in \Re^p, \beta_j = 0, \forall j \in \mathcal{S}^c\} = \beta^{opt} \; a.s. \tag{73}$$

Which is the second result.

□

A.3 Additional lemmas

Lemma 7. *The RE condition in 1 implies (A3) with* $r_{4s} \geq r_e > 0$ *and* $\tilde{p}^* \geq 4s$.

Proof. As in Lemma 1 in Liu et al. (2017). □

Lemma 8. *Let* β^* *be a* S^3ONC *solution to 2 Given (A1) and that* $\mathcal{Q}(\beta^*) \leq \mathcal{Q}(\beta^{true}) + \Gamma$ *holds for some* $\Gamma \geq 0$ *then*

$$\frac{b_l}{2n}\left\|X\delta^*\right\|^2 - \frac{1}{n}W^\intercal X\delta^* \\ \leq \min\left\{\sum_{j\in S} P'_\lambda(|\beta^*_j|)|\beta^{true}_j|,\ \sum_{j\in S} P'_\lambda(|\beta^*_j|)|\beta^*_j - \beta^{true}_j|,\ P_\lambda(a\lambda)(|\mathcal{S}| - \|\beta^*\|_0) + \Gamma\right\}, \quad a.s. \tag{74}$$

Proof. First, we know that $\beta^* \in \arg\min_\beta\{\sum_{i=1}^n \ell(\beta, x_i, y_i) + \sum_{j=1}^p P'_\lambda(|\beta^*|)|\beta_j|\}$ because the KKT conditions are the same as FONC which β^* satisfies. This gives us that $\sum_{i=1}^n \ell(\beta^*, x_i, y_i) + \sum_{j=1}^p P'_\lambda(|\beta^*|)|\beta^*_j| \leq \sum_{i=1}^n \ell(\beta^{true}, x_i, y_i) + \sum_{j=1}^p P'_\lambda(|\beta^*|)|\beta^{true}_j|$. This can be used along the same lines as the level set inequality in the derivation for 8 to get $\frac{b_l}{2n}\left\|X\delta^*\right\|^2 - \frac{1}{n}W^\intercal X\delta^* \leq \sum_{j=1}^p P'_\lambda(\beta^*_j)(|\beta^{true}_j| - |\beta^*_j|)$

The first two terms of the min function are easily obtained from this. The last term can be obtained from 8 by noting that due to Corollary 4, $\beta^* \notin (0, a\lambda)$ and that $P_\lambda(a\lambda) = P_\lambda(\beta) \quad \forall\beta \geq a\lambda$. This gives us that $\frac{b_l}{2n}\left\|X\delta^*\right\|^2 - \frac{1}{n}W^\intercal X\delta^* \leq P_\lambda(a\lambda)(\mathcal{S} - \|\beta^*\|_0) + \Gamma$ Which is the final term to complete the desired result.

□

Lemma 9. *Consider a subgaussian $\tilde{n}$-dimensional random vector $\tilde{W} \in \Re^{\tilde{n}}$ as defined in (A2). Then for any $V \in \Re^{\tilde{n}\times\tilde{n}}$ and $\Sigma_v = V^\intercal V$ then $P[\left\|V\tilde{W}\right\|^2 \leq \sigma^2 Tr(\Sigma_v) + 2\sqrt{Tr(\Sigma_v^2)t} + 2\left\|\Sigma_v\right\| t] \geq 1 - \exp(-t)$ for any $t > 0$ where $Tr(\cdot)$ denotes the trace of a matrix.*

Proof. As in Theorem 2.1 in Hsu et al. (2012). □